

ZEERAK TALAT | LORENTZ WORKSHOP ON INTERSECTIONALITY AND ALGORITHMIC DISCRIMINATION |  
20.12.2017

---

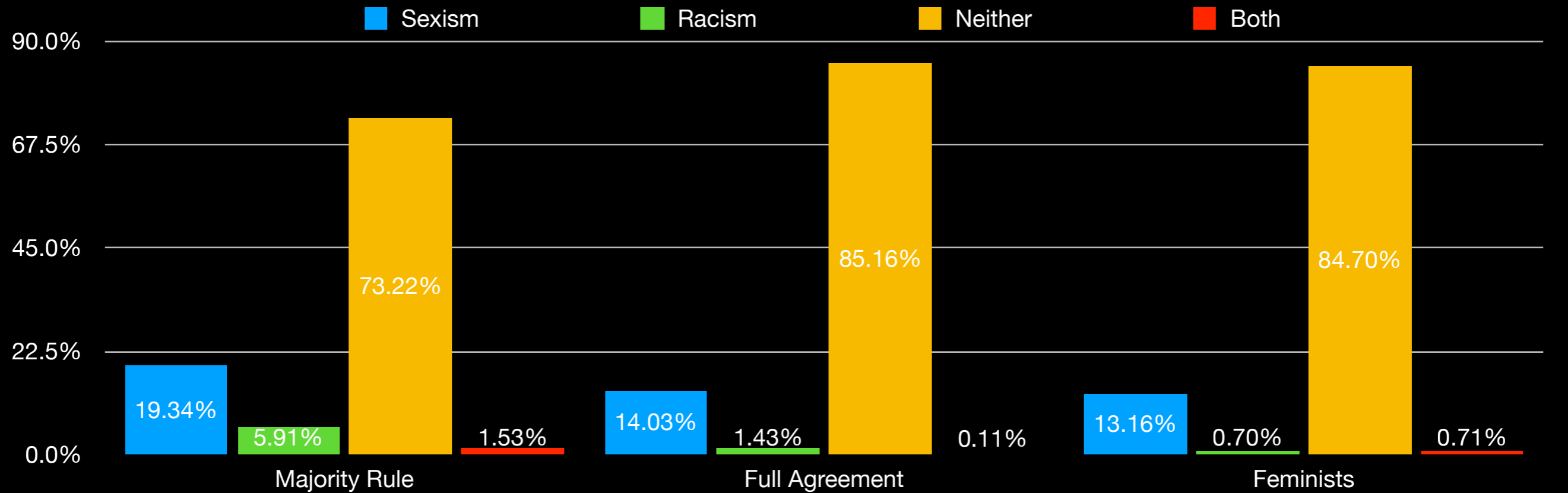
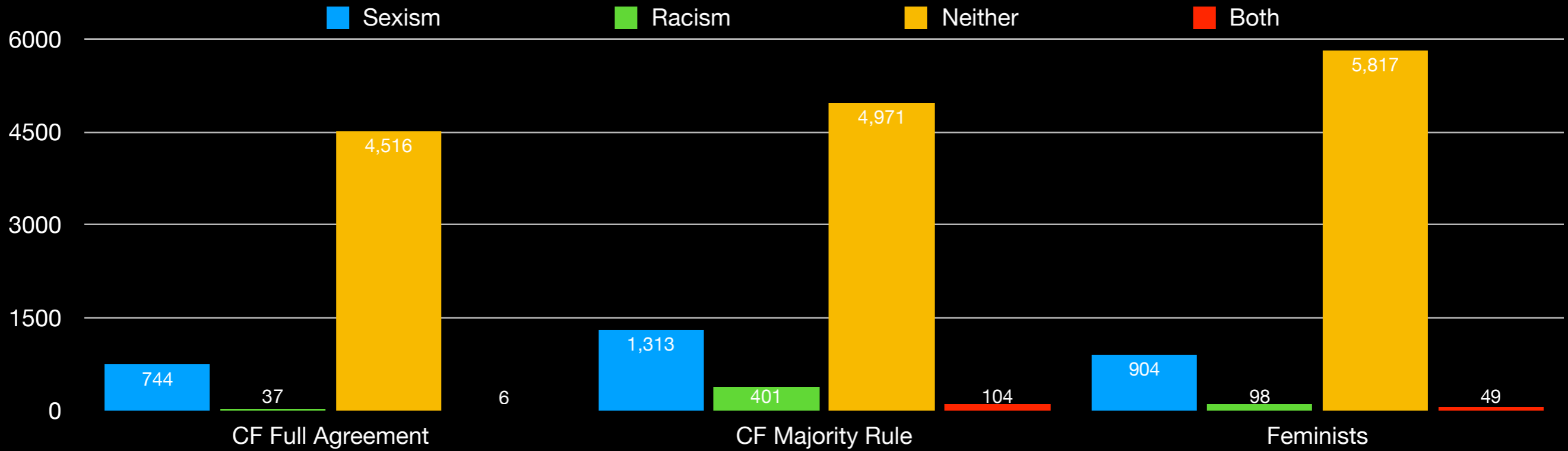
# DETECTING HATE SPEECH USING NATURAL LANGUAGE PROCESSING

## WHY WOULD ANYONE DO THIS?

- ▶ Gamergate
- ▶ The Task
  - ▶ Detecting abuse against women
  - ▶ ... and people of colour
- ▶ The Data
  - ▶ Twitter

# ANNOTATING

1. uses a sexist or racial slur.
2. attacks a minority.
3. seeks to silence a minority.
4. criticizes a minority (without a well founded argument).
5. promotes, but does not directly use, hate speech or violent crime.
6. criticizes a minority and uses a straw man argument.
7. blatantly misrepresents truth or seeks to distort views on a minority with unfounded claims.
8. shows support of problematic hash tags. E.g. “#BanIslam”, “#whoriental”, “#whitegenocide”
9. negatively stereotypes a minority.
10. defends xenophobia or sexism.
11. contains a screen name that is offensive, as per the previous criteria, the tweet is ambiguous (at best), and the tweet is on a topic that satisfies any of the above criteria.



## HOW WELL DO OUR ANNOTATORS AGREE

	Majority v. Feminists	Full Agreement v. Feminists	All CF Annotators
Cohen's kappa	0.34	0.70	0.57
Krippendorff's alpha	0.32	0.70	

## SOME FORMULAS

- ▶  $Precision = \frac{TP}{TP + FP}$
- ▶  $Recall = \frac{TP}{TP + FN}$
- ▶  $F1\text{-score} = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$

## WHAT DOES OUR CLASSIFIER THINK

### F1-Scores

	Amateur (Majority Vote)	Expert/Feminists
<i>Character n-gram</i>	86.41	91.24
<i>Token n-gram</i>	86.37	91.55
<b>Binary Gender</b>	76.64	77.77
<b>GenderProbability</b>	86.37	81.30
<i>Brown Clusters</i>	84.50	87.74
AHST	71.71	55.40

	Explicit	Implicit
Directed	Unambiguous in its potential to be abusive, i.e. use of slurs directed at an individual/entity.	Not immediately clearly abusive. Often obscured by ambiguous terms, sarcasm, lack of profanity, etc. Directed at an entity/individual.
Generalized	Unambiguous in its potential to be abusive, i.e. use of slurs directed at a generalised <i>other</i> .	Not immediately clearly abusive. Often obscured by ambiguous terms, sarcasm, lack of profanity, etc. Directed at an generalised <i>other</i> .



THANK YOU FOR  
YOUR ATTENTION

# SHAMELESS PROMOTION

- ▶ 2nd Workshop on Abusive Language Online @ EMNLP
- ▶ Looking for non-reviewers outside of NLP
- ▶ ... and suggestions for speakers/panelists working with abuse/hate speech from a policy perspective
- ▶ ... and ideas on how we can make people from fields outside of NLP submit

*in Brussels Oct 31st/Nov 1st*



	Explicit	Implicit
Directed	<p>“@User shut yo beaner ass up sp*c and hop your f*ggot ass back across the border little n*gga”</p>	<p>“(((@User))) and what is your job? Writing cuck articles and slurping Google balls? #Dumbgoogles”</p>
Generalized	<p>“So an 11 year old n*gger girl killed herself over my tweets? ^^ thats another n*gger off the streets!!”</p>	<p>“Totally fed up with the way this country has turned into a haven for terrorists. Send them all back home.”</p>